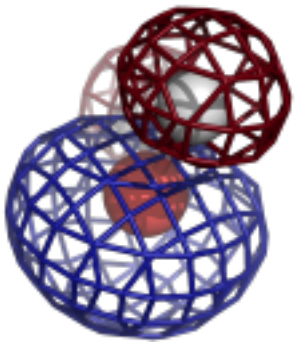
VOLUME SEVEN



COMPUTATIONAL CRYSTALLOGRAPHY NEWSLETTER

JANUARY MMXVI

HOOGSTEEN, BASE PAIRS, VISUALISATION, AMBER

Table of Contents

Editor

Nigel W. Moriarty, NWMoriarty@LBL.Gov

Phenix News

Announcements

Amber

Amber has been added as an option for the chemical information in a macromolecular refinement in Phenix. Starting with dev-2247, the documentation and code is synced to enable the use of Amber in refinement and geometry minimisation. Furthermore, once Amber has been configured, the Phenix GUI adds an input tab with the appropriate program inputs.

AFITT

In a similar fashion, OpenEye's AFITT program for providing chemical information about ligands has been added to Phenix including documentation and GUI interface.

X-ray diffraction indexing and integration programs DIALS and Xia2 now available in Phenix

Two software packages for the analysis and reduction of X-ray diffraction images are now being included in nightly Phenix builds starting at dev-2307: DIALS (Diffraction Integration for Advanced Light Sources) and Xia2. DIALS is a new implementation of spotfinding, indexing and integration algorithms useful for processing x-ray diffraction data prior to scaling and merging. Created as a collaboration between developers at the Diamond Light Source, members of the Computational Crystallographic Initiative at Lawrence Berkeley National Labs, and members of CCP4, the software provides an extensible framework for algorithms needed for diffraction data reduction and visualization. The software is built on the *cctbx* libraries and implements well-established indexing and integration

methods. Furthermore, its extensible design is providing a platform for the creation of new methods, such as new techniques for indexing multiple, overlapping diffraction patterns, new parameterizations for the refinement of crystallographic models, and new treatments for background and signal during integration. Xia2, developed at the Diamond Light Source, is an expert system designed to automate and simplify diffraction data reduction and merging using, if available on the user's system, Mosflm, Scala, XDS, LABELIT, Aimless, Pointless, and recently, DIALS. Note that Xia2 requires CCP4 as a dependency.

Project websites for tutorials and documentation are available at dials.github.io/ and xia2.github.io/

References

Waterman DG, Winter G, Parkhurst JM, Fuentes-Montero L, Hattne J, Brewster A, Sauter NK, Evans G (2013). "The DIALS framework for integration software." *CCP4 Newsletter on Protein Crystallography* **49**, 16-19.

Winter, G (2010). "xia2: an expert system for macromolecular crystallography data reduction." *Journal of Applied Crystallography* **43**, 186-190.

Acknowledgements

DIALS development at Diamond Light Source is supported by the BioStruct-X EU grant, Diamond Light Source, and CCP4.

DIALS development at Lawrence Berkeley National Laboratory is supported by National Institutes of Health / National Institute of General Medical Sciences grant R01-GM095887. Work at LBNL is performed under Department of Energy contract DE-AC02-05CH11231.

New programs

phenix.AmberPrep

Amber requires two additional files (`.prmtop` and `.rst7`) be supplied to a refinement. In addition, the ordering of the input PDB model file must match order in these files. AmberPrep reads a model PDB file and generates a new model PDB and the two additional files for input into refinement. Ligands with the correct code will also be processed.

New features

phenix.structure_search

- Quickly (~1s) find homologous structures for a given model from included internal database. No network is required.
- Option to compile and output a list of ligands found in structures of identified homologs.
- Option to do Blast search for a given model or a sequence file locally. No network is required.
- Option to use a local PDB mirror for PDB file retrieval.

Crystallographic meetings and workshops

Seventh edition of the Macromolecular Crystallography School 2016, May 25-29

Location: Institute Química-Física Rocasolano CSIC (Madrid, Spain).

2016 Annual Meeting of the American Crystallographic Association, July 22-26

Location: Denver, CO.

The 30th European Crystallographic Meeting, August 28-September 1, 2016

Location: Basel, Switzerland.

Expert advice

Fitting Tip #11 – Can a helical DNA base pair be Hoogsteen?

Bradley Hintze and Jane Richardson, *Duke University*

The nice thing about DNA for a crystallographer is that it is nearly always B-form helix, perhaps with interesting perturbations, and that the G•C and A•T base pairs will have the canonical Watson-Crick (WC) geometry. However, very occasionally a Hoogsteen (HG) base pair is identified, usually in a functionally suggestive place (Aishima 2002; Abrescia 2004; Kitayner 2010). Spin-relaxation NMR studies show quite clearly that normal B-

form base pairs in DNA sample the Hoogsteen arrangement transiently, at populations of about 0.5% and lifetimes of about 1 ms (Nikolova 2011; Alvey 2014). In fact, in the first crystal structure of a DNA base pair, Karst Hoogsteen observed an A•T in the non-WC arrangement that now bears his name (Hoogsteen 1959). That arrangement involves a 180° flip of the purine (A or G) base around the glycosidic bond (from *anti* to *syn*), different H-bond partners, and a shortening of the C1'-C1' distance across the helix (see figure 1). This change is harder for a G•C than for an A•T base pair, because the cytosine N3 must be protonated, and the H-bonds are reduced from 3 in WC to 2 in HG.

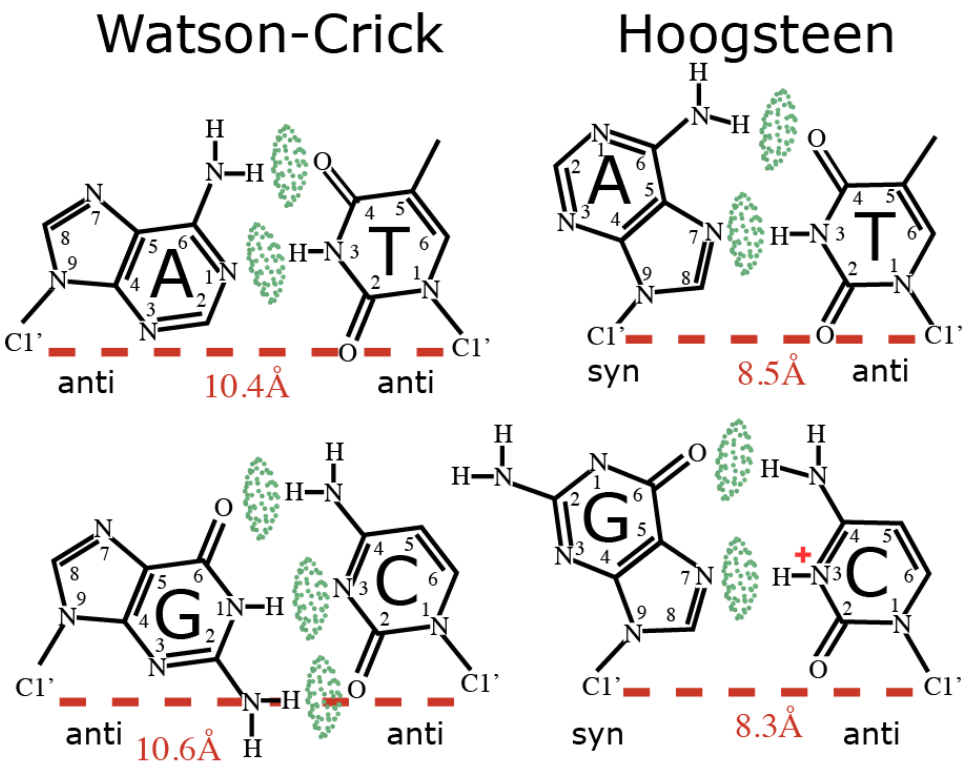


Figure 1: A comparison of the Watson-Crick (WC; left) and Hoogsteen (HG; right) basepairing arrangements. H-bonds are shown by the lens shapes of pale green all-atom contact dots, and the diagnostic distances between C1' atoms are marked in red. Note the plus charge from protonation of the N3 of the cytosine that must happen to enable the less favorable G•C Hoogsteen basepair.

Other contexts for Hoogsteens

Hoogsteen base pairs are a common component of base triples in either RNA or DNA (since after making one Watson-Crick pair the Hoogsteen edge is the next best available for H-bonding). They are not detected, even transiently, in RNA A-form helices, and are now known to be rare but possible in DNA B-form helices.

Occurrence in DNA crystal structures

A recent survey of DNA structures at ≤ 3.0Å resolution found 106 A•T and 34 G•C base pairs that satisfied all three criteria: HG-type H-bonds (each ≤ 3.5Å), *syn* purine χ, and C1'-C1' distance ≤ 9.5Å (Zhou 2014), many of them not mentioned in the accompanying publications. The overall occurrence level of 0.3%, the preference for A•T over G•C, and the preference for TA over AT steps along the sequence, are all quite consistent with the solution NMR and other previous data. New results are an enrichment of HG at helix ends over interiors, a preference for protein or ligand complexes over naked DNA, a preference for the *syn* purine to be at the 5' end vs the 3' end of the helix, and for interior HG pairs to bend the helix slightly toward the major groove side. Although very clear in these crystal structures (e.g. Figure 2) and seen transiently by NMR, there has yet to be NMR evidence of persistent HG base pairs.

Refitting

We have identified a few unambiguous cases, especially in A-tracts, where a base pair modeled as WC but with a shortened C1'-C1' distance can be rebuilt and refined as an HG pair. The new fit has better sterics (H-bonds or clashes), better geometry, and a better fit to the electron density. An example is dA c1 to dT h1 in 3ufd (see Figure 3). Mixtures of WC and HG alternate conformations would also be expected, but of course are harder to verify. We successfully fit such a case for dA 2 of 4auw, where each conformation has a positive difference peak next to the N7, while both conformations together as alternates refine well and lose the difference

density. Note that any WC vs HG rebuild requires significant change in the backbone conformation, especially for the purine.

Conclusion

Our data show that Hoogsteen basepairs are very rare but when they do occur it's mostly in the context of a protein/ligand-DNA complex or at duplex termini. Despite their rarity, it is well worth considering a Hoogsteen basepair when fitting B-form helical DNA. This is especially true if you encounter interpretable difference peaks on either side of the purine, suggestive clashes, or a shortened C1'-C1' distance. As with all refittings, the new fit should improve geometry, stertics, and/or the fit to density. The latter criterion is especially important – Hoogsteen basepairs are extremely rare, so don't do it too often!

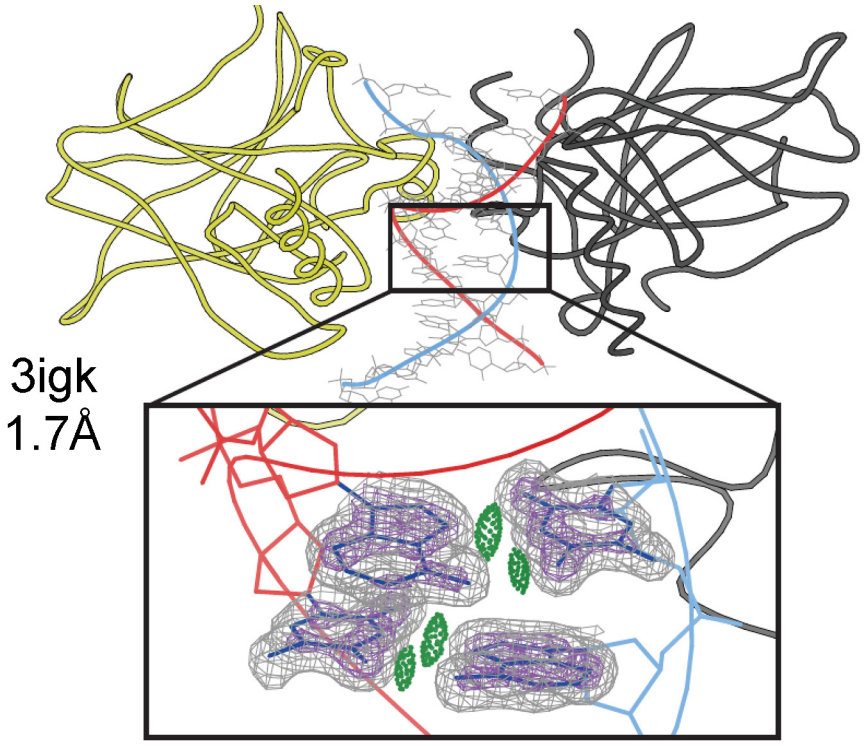


Figure 2: An indisputably correct tandem pair of A•T Hoogsteen base pairs in a complex of p53 protein recognizing an atypical duplex DNA sequence (3igk; Kitayner 2010). From Hintze 2015.

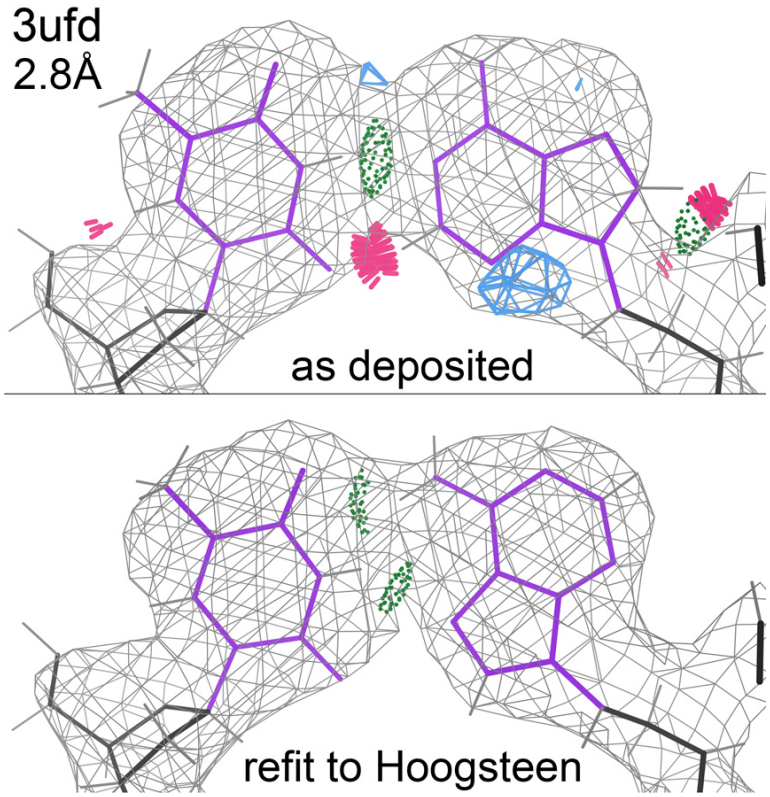


Figure 3: A problematic Watson-Crick base pair at top, with a large clash, distorted base geometry, only one H-bond between the bases, and a positive difference peak (blue) next to the purine N7 (4auw dA h2; Textor 2013). A rebuild that flipped the A base from *anti* to *syn*, adjusted the backbone, and re-refined, produced the much more satisfying fit shown below.

References:

Abrescia NG, Gonzales C, Gouyette C, Subirana JA (2004) "X-ray and NMR studies of the DNA oligomer d(ATATAT): Hoogsteen base pairing in duplex DNA", *Biochemistry* **43**: 4092-4100

Aishima J, Gitti RK, Noah JE, Gan HH, Schlick T, Wolberger C (2002) "A Hoogsteen base pair embedded in undistorted B-DNA", *Nucleic Acids Resl* **30**: 5244-5252

Alvey HS, Gottardo FL, Nikolova EN, Al-Hashimi HM (2014)

"Widespread transient Hoogsteen base pairs in canonical duplex DNA", *Nat Commun* **5**: 4786

Hintze BJ (2015), "Rare Sidechain Conformations in Proteins and DNA," Ph.D. thesis, Duke University.

Hoogsteen K (1959) "The Structure of Crystals Containing a Hydrogen-Bonded Complex of 1-Methylthymine and 9-Methyladenine", *Acta Crystallogr* **12**: 822-823

Kitayner M, Rozenberg H, Rohs R, Suad O, Rabinovich D, Honig B, Shakked Z (2010) "Diversity in DNA recognition by p53 revealed by crystal structures with Hoogsteen base pairs", *Nat Struct Mol Biol* **17**: 423-4292403-240

Nikolova EN, Kim E, Wise AA, O'Brien PJ, Andricoaei I, Al-Hashimi HM (2011) "Transient Hoogsteen base pairs in canonical duplex DNA", *Nature* **470**: 498-502

Textor LC, Holton S, Wilmanns M (2013) "Crystal structure of the bZIP homodimeric Mafb in complex with the C-Mare binding site", unpublished

Zhou H, Hintze BJ, Kimsey IJ, Sathyamoorthy B, Yang S, Richardson JS, Al-Hashimi HM (2015) "New insights into Hoogsteen base pairs in DNA duplexes from a structure-based survey", *Nucleic Acids Resl* **43**: 3420-3433

FAQ

Why can't I see the link I want in my model?

Phenix recently (1.10.1-2155) began writing LINK records into the output model file to facilitate better visualisation of the model used internally by the refinement package including the automatic linking algorithm's results. The second short communication has a discussion of the relationship between a refinement package and a visualisation package.

Phenix, in general, does not use the LINK records in the input model.

Characterization of base pair geometry

Xiang-Jun Lu,[a] & Wilma K. Olson[b]
[a]*Department of Biological Sciences, Columbia University, New York, NY 10027*
[b]*Department of Chemistry and Chemical Biology, Rutgers – The State University of New Jersey, Piscataway, NJ 08854*

Correspondence email: xiangjun@x3dna.org

Introduction

The interactions of the planar nitrogenous bases (A, C, G, T, and U) play a critical role in the three-dimensional organization of DNA and RNA. The relative spatial arrangement of a pair of associated bases can be rigorously quantified by six rigid-body parameters (Dickerson et al., 1989): three translational parameters called Shear, Stretch, and Stagger, and three rotational parameters denoted Buckle, Propeller (twist), and Opening (figure 1). The numerical values of these base pair parameters or the six step parameters used to describe the positioning of neighboring base pairs depend upon the choice of reference frame (Lu et al., 1999). The establishment of a standard base reference frame (Olson et al., 2001) and its implementation in 3DNA (Lu et al., 2003) and Curves+ (Lavery et al., 2009) has largely resolved discrepancies in the analysis of double-helical structures of Watson-Crick (WC) base pair.

The standard reference frame (Olson et al., 2001) is base centric, defined with respect to an 'idealized', perfectly planar WC pair where all six parameters are null. As noted in the classic Dickerson B-DNA dodecamer (Drew et al., 1981), WC pairs are normally non-planar, with Propeller (and Buckle) significantly different from zero. Notably, Propeller in right-handed DNA double helices has a mean value of around -12° (Dickerson et al., 1989; Olson et al., 2001), and its persistence has been rationalized in terms of the increased base-stacking interactions found in these structures (Calladine, 1982; Levitt, 1978). More generally, among the six base pair parameters, Buckle, Propeller, and Stagger describe the *non-planarity* of a pair: Buckle and Propeller render the two bases *non-parallel*, whilst Stagger leads to a vertical separation. On the other hand, Shear, Stretch, and Opening are critical for characterizing different types of non-WC pairs (Lu et al., 2015; Lu et al., 2003). For example, the G–U wobble pair is characterized by a Shear of -2.2 Å; if counted the other way around, i.e., as U–G, the Shear value is +2.2 Å instead.

3DNA (Lu et al., 2003) uses the standard base reference frame (Olson et al., 2001) to calculate six *local* base pair parameters with its 'analyze' routine. The parameters unambiguously characterize any base pair, be it WC, G–U wobble, or non-canonical (e.g., a Hoogsteen pair). Conversely, given the six parameters and base sequence, the 3DNA 'rebuild' routine rigorously reconstructs the spatial disposition of the two interacting bases. This reversibility is one of the unique features of 3DNA, applicable to pairs of bases in both DNA and RNA.

Simple base pair parameters

Although the six *local* base pair parameters are rigorous and serve a good purpose, their numerical values can be cryptic for non-canonical pairs, most notably the values of Buckle and Propeller. Two recent CCN articles (Richardson, 2015a, 2015b) emphasized the importance of base

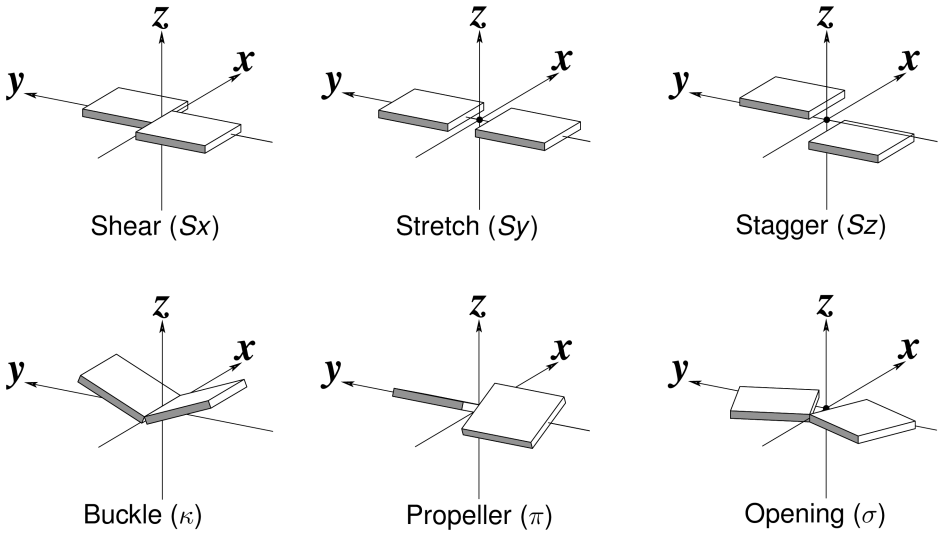


Figure 1: Schematic diagrams of the six rigid-body parameters commonly used for the characterization of base pair geometry.

pair non-planarity at biologically significant positions in high-resolution nucleic acid structures (e.g., functional binding sites) and the need to account correctly for this non-planarity in deriving models of DNA and RNA based on low-resolution data. The account of two of the classic measures of base pair deformations (figure 1) given in the article — "the 'propeller-twist' torsion around a line joining the two bases, and the 'buckle' angle of their bend across the line of base pair H-bonds" — prompted us to derive a new set of six *simple* parameters for a complete qualitative description of base pair geometry. The term '*Simple*' is used because the parameters are more intuitive for non-canonical pairs (figure 2), and to differentiate them from the existing *local* base pair parameters. The *simple* parameters have been implemented in both the 3DNA 'analyze' routine and the new DSSR software (Lu et al., 2015).

Detailed definitions of the *simple* base pair parameters, with worked examples, can be found on the 3DNA homepage (x3dna.org). The key differences between the *simple* and *local* parameters lie in the definition of the base pair coordinate frame and in the description of angular parameters. The *simple* treatment uses the locations of atoms on the interacting purine (R) and pyrimidine (Y) bases, either RC8/YC6 (default) or RN9/YN1, as the (long) ***y***-axis of the pair, corrected to be orthogonal with the ***z***-axis. The ***z***-axis is the average of the *z*-axes of the two bases, following the definition of the corresponding *local* base pair axis and taking the anti-parallel direction into consideration (e.g., in WC pairs). The (short) ***x***-axis fulfills the right-handed rule. The three translational parameters (figure 1) are simply the projections of the vector linking the origins of the two base-reference frames onto the respective ***x***-, ***y***- and ***z***-axes, similar to the definition of the corresponding *local* base pair parameters. The three *simple* angular parameters, however, are defined as 'torsion' angles: Buckle as the 'torsion' angle of the *z*-axes of the two bases with respect to the short base pair axis, Propeller as the 'torsion' angle of the two *z*-axes with respect to the long base pair axis, and Opening as the 'torsion' angle of the two *y*-axes with respect to the base pair ***z***-axis. While the *simple* base pair parameters of WC pairs closely

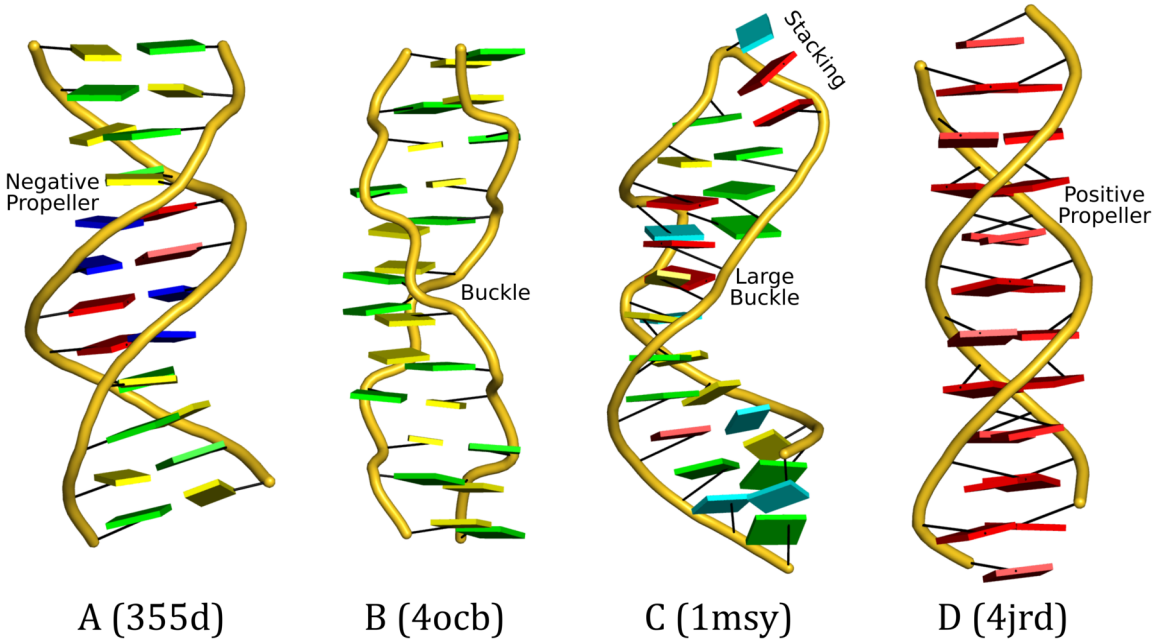


Figure 2: DSSR-introduced cartoon-block representations of DNA and RNA structures that combine PyMOL cartoon schematics with color-coded rectangular base blocks: A, red; C, yellow; G, green; T, blue; and U, cyan. (A) The Dickerson B-DNA dodecamer solved at 1.4-Å resolution [PDB id: 355d (Shui et al., 1998)], with significant negative Propeller. (B) The Z-DNA dodecamer [PDB id: 4ocb (Luo et al., 2014)], with virtually co-planar C–G pairs at the ends, and noticeable Buckle in the middle. (C) The GUAA tetraloop mutant of the sarcin/ricin domain from *E. coli* 23 S rRNA [PDB id: 1msy (Correll et al., 2003)], with large Buckle in the A+C pair, and base-stacking interactions of UAA in the GUAA tetraloop (upper-right corner). (D) The parallel double-stranded poly(A) RNA helix [PDB id: 4jrd (Safaee et al., 2013)], with up to +14° Propeller. The simple, informative cartoon-block representations facilitate understanding of the base interactions in small to mid-sized nucleic acid structures like these. The base identity, pairing geometry, and stacking interactions are obvious.

resemble the *local* base pair parameters, the two sets can differ significantly for non-canonical pairs.

Cartoon-block representations

The new DSSR component of 3DNA (Lu et al., 2015) includes an easy way to create highly effective cartoon-block representations that showcase base pair geometry and base-stacking interactions (figure 2). DSSR has been tested against all nucleic-acid-containing structures in the Protein Data Bank (PDB), and works with atomic coordinate files in either PDB or PDBx/mmCIF format. Figure 2 illustrates the arrangements of bases in four representative high-resolution X-ray crystal structures. The following command generates the script, named '`355d.pml`', needed to construct figure 2A (the Dickerson B-DNA dodecamer solved at 1.4 Å resolution, PDB id: 355d)within PyMOL:

```
x3dna-dssr -i=355d.pdb -o=355d.pml\
    --cartoon-block
```

Once '`355d.pml`' is loaded into PyMOL, the commands '`orient; turn z, -90; ray`' set the structure in the most extended view vertically and perform the ray-tracing needed to obtain the high-resolution PNG image. Similar commands apply to the other three cases. Details on reproducing the illustrated images, and outputs (including the *simple* base pair parameters) of DSSR and the 3DNA 'analyze' program for the four structures, are available at: x3dna.org/highlights/CCN-on-base-pair-geometry.

In addition to the default style shown in figure 2, DSSR provides several variations of the cartoon-block representation of planar, aromatic bases. For example, one can change both the thickness and the sizes of the blocks, shade the minor-groove edges of the bases, and represent a WC pair as a single long rectangular base pair block. One can also orient a structure in a specific base or base pair reference frame for easy comparison, and can attach the helical axes to an array of stacked base pairs. See the x3dna.org for more examples.

Summary

Base pair geometry can be described in different ways. The existing set of six *local* base pair parameters in 3DNA is mathematically rigorous, allowing for an unambiguous characterization of any pair of interacting bases and serving as input for exact model building. The new set of six *simple* base pair parameters described herein provides a more intuitive interpretation of intra-base pair structural variations, especially for the out-of-plane Buckle and Propeller distortions of non-canonical base pairs. The two sets of base pair parameters complement one another and serve different audiences and/or purposes. Numerical values of both sets of parameters are readily obtained with 3DNA. Moreover, the DSSR component of 3DNA provides highly effective cartoon-block representations of the base interactions within a nucleic acid structure.

References

Calladine, CR. (1982). Mechanics of sequence-dependent stacking of bases in B-DNA. *J Mol Biol, 161*(2), 343-352.

Correll, CC, & Swinger, K. (2003). Common and distinctive features of GNRA tetraloops based on a GUAA tetraloop structure at 1.4 Å resolution. *RNA, 9*(3), 355-363.

Dickerson, RE, Bansal, M, Calladine, CR, Diekmann, S, Hunter, WN, Kennard, O, von Kitzing, E, Lavery, R, Nelson, HCM, Olson, WK, Saenger, W, Shakked, Z, Sklenar, H, Soumpasis, DM, Tung, C-S, Wang, AH-J, & Zhurkin, VB. (1989). Definitions and nomenclature of nucleic acid structure parameters. *EMBO J, 8*(1), 1-4.

Drew, HR, Wing, RM, Takano, T, Broka, C, Tanaka, S, Itakura, K, & Dickerson, RE. (1981). Structure of a B-DNA dodecamer: conformation and dynamics. *Proc Natl Acad Sci U S A, 78*(4), 2179-2183.

Lavery, R, Moakher, M, Maddocks, JH, Petkeviciute, D, & Zakrzewska, K. (2009). Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res, 37*(17), 5917-5929.

Levitt, M. (1978). How many base pairs per turn does DNA have in solution and in chromatin? Some theoretical calculations. *Proc Natl Acad Sci U S A, 75*(2), 640-644.

Lu, XJ, Bussemaker, HJ, & Olson, WK. (2015). DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res, 43*(21), e142.

Lu, XJ, & Olson, WK. (1999). Resolving the discrepancies among nucleic acid conformational analyses. *J Mol Biol, 285*(4), 1563-1575.

Lu, XJ, & Olson, WK. (2003). 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res, 31*(17), 5108-5121.

Luo, Z, Dauter, M, & Dauter, Z. (2014). Phosphates in the Z-DNA dodecamer are flexible, but their P-SAD signal is sufficient for structure solution. *Acta Crystallogr D Biol Crystallogr, 70*(Pt 7), 1790-1800.

Olson, WK, Bansal, M, Burley, SK, Dickerson, RE, Gerstein, M, Harvey, SC, Heinemann, U, Lu, XJ, Neidle, S, Shakked, Z, Sklenar, H, Suzuki, M, Tung, CS, Westhof, E, Wolberger, C, & Berman, HM. (2001). A standard reference frame for the description of nucleic acid base pair geometry. *J Mol Biol, 313*(1), 229-237.

Richardson, JS. (2015a). A context-sensitive guide to RNA & DNA base pair & base-stack geometry. *Computational Crystallography Newsletter, 6*(Part 2), 47-53.

Richardson, JS. (2015b). Fitting Tip #10 – How do your base pairs touch and twist? *Computational Crystallography Newsletter, 6*(Part 2), 28-31.

Safaee, N, Noronha, AM, Rodionov, D, Kozlov, G, Wilds, CJ, Sheldrick, GM, & Gehring, K. (2013). Structure of the parallel duplex of poly(A) RNA: evaluation of a 50 year-old prediction. *Angew Chem Int Ed Engl, 52*(39), 10370-10373.

Shui, X, McFail-Isom, L, Hu, GG, & Williams, LD. (1998). The B-DNA dodecamer at high resolution reveals a spine of water on sodium. *Biochemistry, 37*(23), 8341-8355.

Note added in proof

The DSSR cartoon-block representation (figure 2) has been integrated into PyMOL via the 'dssr_block' command (http://pymolwiki.org/index.php/Dssr_block).

Visual representations of internal coordinate restraints: Advantages and limitations

Pavel V. Afonine and Nigel W. Moriarty

Department of Biophysics and Integrated Bioimaging, Lawrence Berkeley Laboratory, Berkeley, CA 94720

Correspondence email: NWMoriarty@LBL.Gov

Macromolecular structure refinement can be thought of as an optimization problem with the goal to obtain a model that describes the experimental data as well as possible. Wikipedia defines optimization as:

> In the simplest case, an optimization problem consists of maximizing or minimizing a real function by systematically choosing input values from within an allowed set and computing the value of the function. The generalization of optimization theory and techniques to other formulations comprises a large area of applied mathematics. More generally, optimization includes finding "best available" values of some objective function given a defined domain (or a set of constraints), including a variety of different types of objective functions and different types of domains.

In case of protein crystallography, there is a model typically consisting of two components: atomic model (protein, DNA/RNA, ordered solvent and ligands) and non-atomic model, such as bulk solvent. There is also experimental data consisting of intensities of diffracted light (X-ray or neutron, for instance). The model is changed systematically to better fit the experimental data.

In other fields such as cryo-electron microscopy (cryo-EM), the model is usually atomic model as well and the data are the 3D reconstruction volume map. Various methods can be used for optimization of the model into the map: from rigid-body docking and flexible fitting to gradient-driven minimization and simulated annealing.

What's common to macro-molecular crystallography and cryo-EM is that experimental data is almost always of insufficient quality to refine parameters of atomic model (coordinates, for example) individually. Therefore to make refinement practical restraints or constraints are used with the corresponding refinements called restrained or constrained refinement. In what follows we focus on restrained refinement. In restrained refinement the target function contains two components

$$T = T_{data} + \omega * T_{restraints}$$

with T_{data} describing how well the model fits the data and $T_{restraints}$ being a source of extra *a priori* knowledge about the molecule added with some weight. This knowledge is the information about covalent bond lengths and angles, dihedral angles, planar molecules (such as phenylalanine ring), chiral centers and nonbonded interactions:

$$T_{restraints} = T_{bond} + T_{angle} + T_{dihedral} + T_{plane} + T_{chiral} + T_{nonbonded}$$

Sometimes even this information isn't sufficient and more is needed. This may be, for example, information about secondary structure arrangements of proteins or nucleic acid molecules, distribution of main chain conformations (Ramachandran plot) or side chain conformations (rotamers). Similar to above, this information is added as extra terms to $T_{restraints}$.

Typically, terms in $T_{restraints}$ are sums of squared differences between values found in current model and "library" values, for

example:

$$T_{bond} = \sum_{\text{pairs of covalently bound atoms}} \frac{1}{\sigma^2}\left(d_{model}-d_{library}\right)^2$$

The library values are tabulated values collected from various sources such as derived from high-quality high-resolution small molecules, spectroscopy experiments and theoretical calculations that are compiled into libraries such as CCP4 Monomer Library (Vagin et al. 2004) or GeoStd (see Notes). Recent developments in restraint libraries include dynamic restraints based on the conformation of the protein backbone (Moriarty et al. 2014).

Refinement packages routinely apply the restraints from the libraries and, in addition, will automatically apply a set of links to provide a more complete chemical picture of the macromolecule. Examples are the peptide link and disulfide bridge. The former is generally based in the order of the protein residues and other criteria such as distance, residue type and/or atom names. The disulfide link is generally based on proximity. In either case a standard link can be applied from the library to create the bonding although Phenix now uses new link information that includes handedness (Sobolev et al. 2015). The library of standard links extends to a subset of the possible carbohydrate links, either protein-carbohydrate or intra-saccharide.

It is not uncommon that a structure may contain novel ligands that are covalently linked to the macromolecule or each other. The novel ligands will often require restraints be generated to add to the restraints target function. It is not unreasonable to expect that the covalent links be non-standard and unique enough to not be present in existing libraries or outside the scope of automatic linking procedures. This means that in order to make a refinement program aware of such unusual bonds (so that the program adds corresponding term to T_{bond}) a user needs to convey the program such information. This can be done by a variety of ways such as using atom selection syntax to specify custom bonds; or creating a link in the same format as the library and specifically applying the link, which allows for more complexity.

A feature of many of the Phenix programs is the ability to write a file that contains all the restraint information used by the program. To ascertain the specified link was actually used in refinement one can inspect `.geo` file that `phenix.refine` creates. The geometry restraints are grouped into restraint types. A typical bond section is shown in schema 1. Each bond is listed in order of decreasing residual and contains the ID string of each atom, the ideal bond length from the library, the value of the bond in the model, the

```
Bond restraints: 4714
Sorted by residual:
bond pdb=" C2D HEM A 201 "
     pdb=" C3D HEM A 201 "
  ideal  model  delta    sigma   weight residual
  1.334  1.521 -0.187 2.00e-02 2.50e+03 8.74e+01
bond pdb=" CA  VAL A  14 "
     pdb=" CB  VAL A  14 "
  ideal  model  delta    sigma   weight residual
  1.537  1.563 -0.025 1.29e-02 6.01e+03 3.88e+00
```

Schema 1: The first three bond restraints of a typical geometry file including a ligand (HEM). Each bond lists the atom ID strings, ideal and actual values, the difference, sigma, weight and residual.

difference of the ideal and actual, the sigma, weight and residual or contribution to T_{bond}. The file is written at the beginning of a refinement and can optionally be written at the end of a refinement, which is useful to check how close the model approaches the restraints. A simple program is available, `elbow.refine_geo_display` that can be used to display restraints of an atom selection.

Visualisation of the refinement model is a powerful tool allowing greater understanding of the model compared to a list of restraints in a `.geo` disk file. However, it is the file that contains the absolute information. When displaying molecules using graphic programs (such as Coot (Emsley et al. 2010) or PyMOL (DeLano 2002)) representations of the bonds in a model are drawn on the screen based on a number of different criteria. The criteria differ between programs and version. Links between entities such as carbohydrate units, novel ligands and metals are especially difficult to represent as the graphics program does not know what the refinement package did internally and can only rely on the information in the model file. These links may not be drawn by the graphic program as a bond (solid or dashed stick connecting two atoms involved), which may result in confusion and suspicion that the bond in question was not used in refinement.

For example, the linking of NAG to ASN is a common glycosidic bond in proteins. When displaying this region of the protein Coot uses the LINK record in the PDB to draw the glycosidic bond between the ASN and NAG (see figure 1a) using a dashed line. Removing the LINK record results in the visualization of the bond being absent (see figure 1b). As stated earlier, the sequence of the entities can lead to linking. In the case of Coot, a PDB with just the ASN and NAG entities results in a solid line representing the bond between the two. Furthermore, changing the code of NAG to LIG and inserting a TER card between the two entities gives the visualisation in figure 1c — a

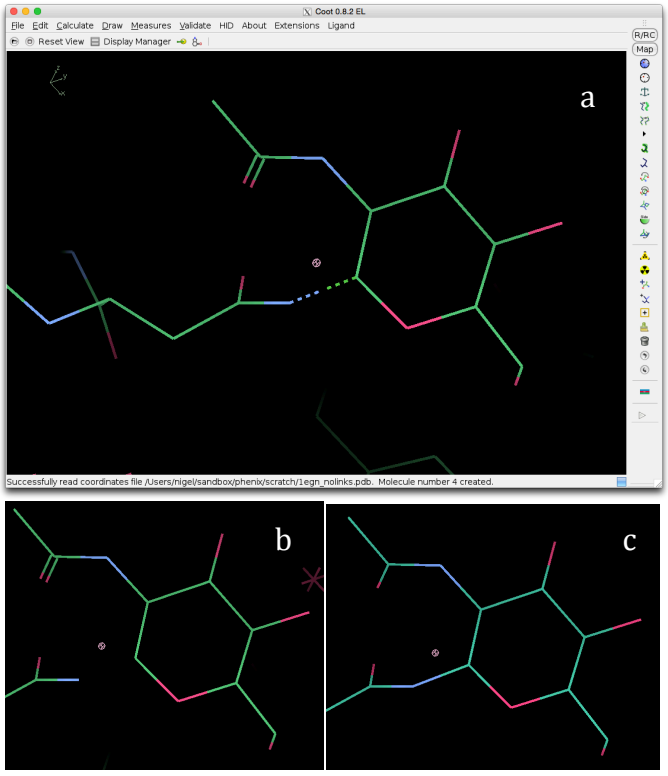


Figure 1: Comparison of the visualization of a glycosidic bond between ASN and NAG with LINK record (a), without LINK record (b) and as a pair of sequential entities (c).

solid line drawn by Coot between the two closet atoms in sequential entities. The visualisation of the ASN-NAG link is unchanged between the entities regardless of whether the link is being specific explicity specified (LINK) or implied (NAG is usually linked to ASN) or explicitly excluded (TER card).

When visualising the same model in PyMOL 0.99, the link is based on distance and ignores completely the identity of the entities. This means that in all cases spoken about in the previous paragraph, there is a solid line between the ASN and NAG representing the glycosidic bond (see figure 2). It should be noted that version 0.99 is quite old and the newer version may behave differently.

Displaying bond valence is also an important part of the visualisation of the model. PyMOL

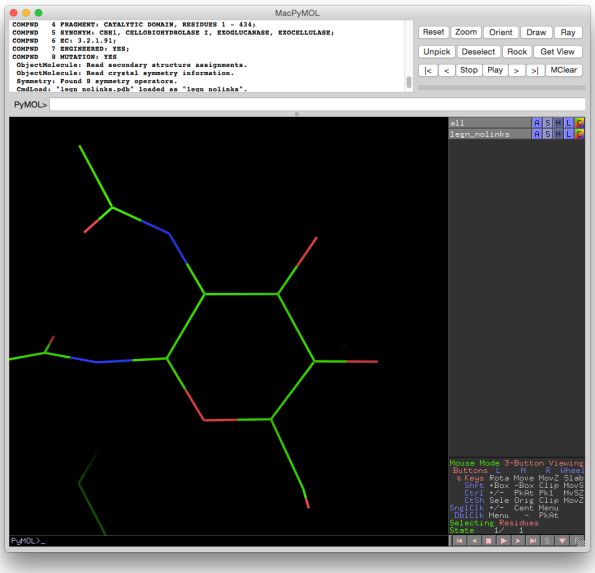


Figure 2: PyMOL visualisation of the NAG in 1EGN.

has a setting that defaults to false. To see the double bond in PyMOL, the command "`set valence, 1`" is required to enable it. The determination of the bond does not rely on the residue code. In the ASN-NAG example, changing the NAG code to LIG does not change the depiction of the double bond as two lines.

An example of when bond valence visualisation can be helpful is p-coumaroyl-shikimate bound to the PvHCT2a protein and deposited in the PDB as 5FAL. There were some important questions that revolved around the bond valence of a particle ring in the ligand. The ring in question (see figure 3) has a single double bond and two chiral centres. The location of the double bond is important for the chemistry of the ligand and mechanism as well as the generation of

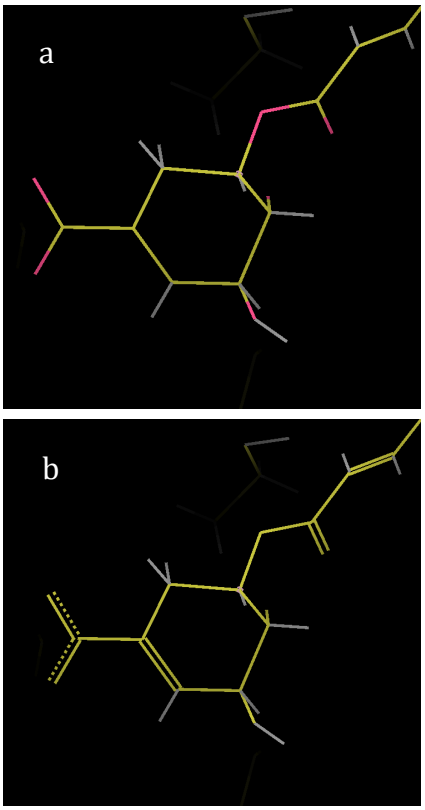


Figure 3: Comparison of a Coot visualization of a novel ligand without the restraints library loaded (a) and with the restraints library loaded (b).

restraints. If the hydrogens had been absent, the visualisation in figure 3a would have been information poor. Loading the novel ligands restraints library into Coot provides the view in figure 3b that is much more informative.

Quick Summary

Visualisation packages do not always know what a refinement package is doing internally and the representation of a model is limited by the data exchange medium.

Notes

GeoStd, An open-source restraints library, http://sourceforge.net/projects/geostd

Wikipedia, The Free Encyclopedia, s.v. "Mathematical optimization" (accessed 26th January, 2016), http://en.wikipedia.org/wiki/Mathematical_optimization

References

DeLano, Warren L. 2002. *PyMOL 0.99*.

Emsley, P., B. Lohkamp, W. G. Scott, and K. Cowtan. 2010. "Features and Development of Coot." Acta Cryst. D 66: 486–501.

Moriarty, Nigel W., Dale E. Tronrud, Paul D. Adams, and P. Andrew Karplus. 2014. "Conformation-Dependent Backbone Geometry Restraints Set a New Standard for Protein Crystallographic Refinement." *FEBS Journal* 281 (18): 4061–71. doi:10.1111/febs.12860.

Sobolev, Oleg V., Nigel W. Moriarty, Pavel V. Afonine, Bradley J. Hintze, David C. Richardson, Jane S. Richardson, and Adams, Paul D. 2015. "Disulfide Bond Restraints." *Computational Crystallography Newsletter* 6 (1): 13–13.

Vagin, A. A., R. A. Steiner, A. A. Lebedev, L. Potterton, S. McNicholas, F. Long, and G. N. Murshudov. 2004. "REFMAC5 Dictionary: Organization of Prior Chemical Knowledge and Guidelines for Its Use." *Acta Crystallographica Section D-Biological Crystallography* 60 (December): 2184–95. doi:10.1107/S0907444904023510.